**Car Price Prediction**

CSDA 5110 Analytics Programming With R  
Tejaswi Amujala

Webster University  
March 15, 2024

**Car Price Prediction**

**Abstract**

In today’s world, life becomes easy with the usage of technology. At the same time it is very important how to use the technology as slight change in data causes a huge loss to large organizations and individuals. Here as I introduce my dataset called “CarPrediction”. In simple words, I want to show how the price of the cars works for different features using Machine learning models such as Multi Linear Regression, Decision Tree and Random forest Models.

**Part 1- Introduction and context**

1. **Introduce your topic and explain the overall situation/landscape.**

In business terms, It is very important to understand the market financial gains and loss. In automobile industry. Because of Machine learning models, Businesses are able to predict the needs and make a deal out of that. In my dataset, the cars which I am predicting the price is for refurbished cars. There are many features such as mileage, make, model, year, condition of the car.

**The Problem**

Basically every buyer and seller first understands the market value of the car price. In olden days, people buy car relying on the preferences of people and historical trends which did not had the capacity to understand the other features such mileage, model,make,year,etc which actually affected their money that is price.

**The solution**

As the generations passed, here comes a new AI-ML Driven generation. With the help of Multi Linear Regression, Decision Tree and Random Forest Models, I am able to predict the price of the car with different types of features/variables in my dataset. My dataset hold 1000 observations, It means the models are able to understand and analyse my large dataset, understand the difficult patterns and give correct accurate results. This is very useful for the automobile market and upcoming various industries.

**why either are important**

This is important in playing Equal fair market pricing, sales and profitability. Here, we understand that there are no more cheatings in price terms as in olden days people seller charges more money and fools people. As we got introduced with models, the pricing is fair and depending on sales it is a great benefit for the automobile industry. Also, there are huge profits. Overall, transparency is important to seller and buyer to make a deal.

1. **Introduce your problem/opportunity statement**

***Understanding the nature challenges-***

1. ***Good or Bad Thing –*** This is very challenging because everything is lacking inbuilt judgement. It may be good or bad. In olden days, analysing the car prices may be correctly accurate or wrong.
2. ***Mitigation or encouragement-*** Basically everything lies on reduction of severity that is mitigation because after the introduction of models it becomes a fair pricing of the car with crystal clear transparency and relevant information.

***Severity or impact if it is not mitigated or encouraged***

If nothing is changed , it really impacts a lot for the generations.

1. **Seller Problems-**

**Issue**- It becomes difficult for them to sell for a right price.

**Result**- Sellers lose their benefits such as money and sales.

1. **Buyer problems***-*

**Issue**- Buyers might feel fooled because they pay too much price without actual judgement*.*

**Result***-* They will feel not to buy cars and lose interest in market or trust on brands.

So overall , we should care for the future markets and individuals.

***Why this is considered or addressed***

The understanding of accepting the accurate pricing is extremely important as it gives equal pricing between seller and buyer which is considered to create a trust towards the company, the clear clarity of the market trends. Mainly buyer who is confident on the pricing of the car brings more sales to the company which maintains a healthy business and relationship with the customers.

1. **What will you prove in this paper?**
   1. ***Tell me what it is that you end up finding-***

In my project, I ended up analyzing my models such as Multi linear Regression, Decision Tree and Random Forest Models in which “Multi Linear Regression Model” shows best fit. As my target variable is Price for all three models, It shows a lots of differences using these models.

* 1. ***Tell me how you found it-***

I found these by evaluation metrics such as accuracy and error scores/measures. I used Mean Absolute error,Root Mean Squared Error and R-Squared and compared. Here, I see multi linear regression values are best fit.

1. MAE and RMSE values are low compare with other models.
2. R Squared value is high compare with other models.
   1. ***Tell me why your dataset is able to prove whatever it is you are presenting***-

My dataset represents the quality of my target variable “Price”. It portrays the actual car price in the market and represents correct measurement. Also, selection of models has the power to predict the results.

**Part 2- Exploratory Analysis of Data and variables**

1. **Variable analysis-**

*Data pulling-*

* My dataset was pulled from Kaggle website and downloaded.
* I opened R-Studio and import the excel data, got the code and ran it.
* I named my dataset as “CarPrediction”

1. **Variable means, what it represents, and how it is measured**-

There are total 6 variables in my dataset. Make, Model, Condition, Price, Year, Mileage.

Price is my target variable. Let me explain each variable step by step.

**Mileage**:

*Meaning*:- The Mileage explains overall distance the car drives

*Representation*: -It is a Numeric value represents miles.

*Measurement*:- The measurement is in miles and it stands on continuous scale

**Year**:

*Meaning*:- The Year explains the year which the car is introduced or manufactured.

*Representation*:- It is a Numeric value represents the car introduction year

*Measurement*:- The measurement is in years and it stands on discrete scale

**Make**:

*Meaning*:- The Make explains the company or brand of car.

*Representation*:- It is a Categorical value represents which brand it is

*Measurement*:- There is no measurement as it is categorical value and it stands on nominal scale.

**Model**:

*Meaning*:-The Model explains the sub model or style of the car.

*Representation*:- The Categorical values represents the style of the car

*Measurement*: -There is no measurement as it is categorical value and it stands on nominal scale.

**Condition**:

*Meaning*:- The Condition represents the performance of the car

*Representation*:- It is a Categorical values represents the condition/performance of the car. Good , excellent like that.

*Measurement*: -There is no measurement as it is categorical value and it stands on ordinal scale.

**Price**:

*Meaning*:-The Price represents the value of the car(dollars money)

*Representation*:- It is Numeric values indicating value of the car.

*Measurement*:- It is measured in price cash currency ( dollars).

1. **Discuss the range or whether it is positively or negatively correlated to the response variable, if numeric**-

There are 3 numeric variables**.**

**Mileage**:

*Range*: It portrays range in which car drove.

*Correlation*: Mileage is negatively correlated with the response variable (Price) .Basically, More the mileage goes up high, the less is the value of the cars. As people who decide to buy a car wants to get the car with low mileage.

**Year**:

*Range*: It portrays the range of car introduction of the year from oldest to newest years.

*Correlation*: Basically, Year is positively correlated with the response variable(price). If the car is of latest model, buyers wants to buy it because of new features and sophisticated technology.

**Price**:

*Range*: It portrays the range of cost of the car values.

*Correlation*: Being price as the response variable, it has no correlation. It correlates with other numeric variables.

1. **If it is a categorical variable, describe the different levels and what they mean-**

**Make:**

*Levels*: It describes the style or brand of car like ford , toyota etc

*Meaning*: It is a categorical variable and based on brands of the car.

**Model:**

*Levels*: various car models like Civic, Accord, etc

*Meaning*: It represents styles or versions of the car.

**Condition:**

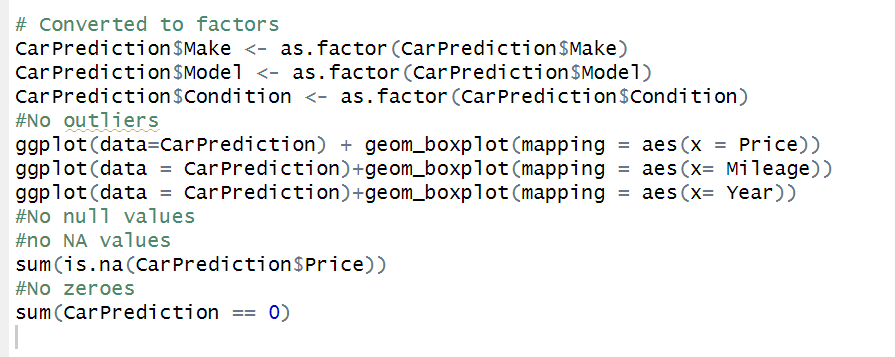
*Levels*: Quality of the car. For example, it says in which condition is it like fair, good, excellent

*Meaning*: There are 3 conditions . Excellent shows no issue with car, fair shows their might be issue with car small issues.

1. **Discuss outliers, null values, zeros, and how you addressed-**

After detailed analyzing of my dataset, I did not find any outliers, null and zero values in my dataset.

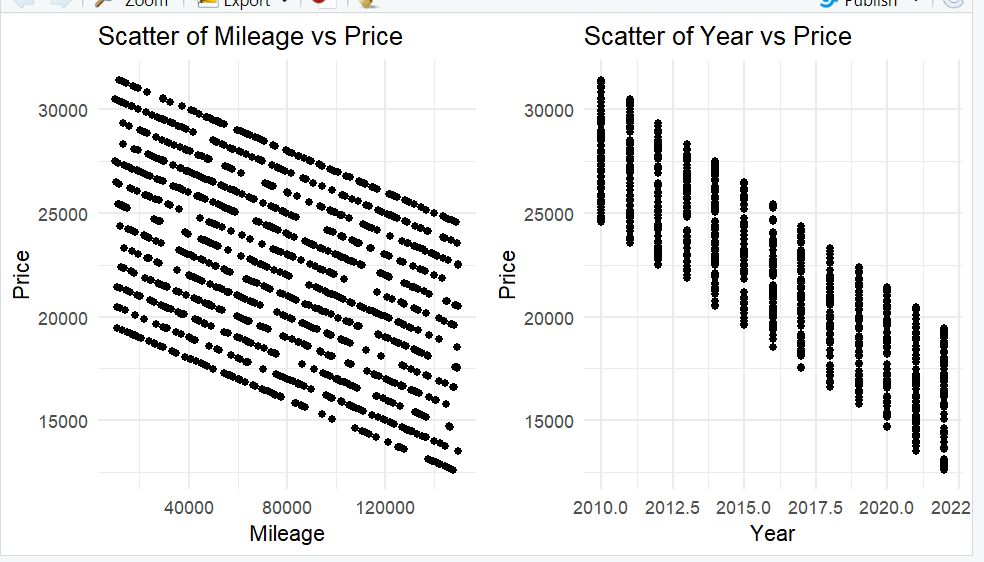
Addressed using these codes-



1. **Analyze variables and their relationship to your response variable and/or each other –**

Create graph shows its relationship to the response variable-

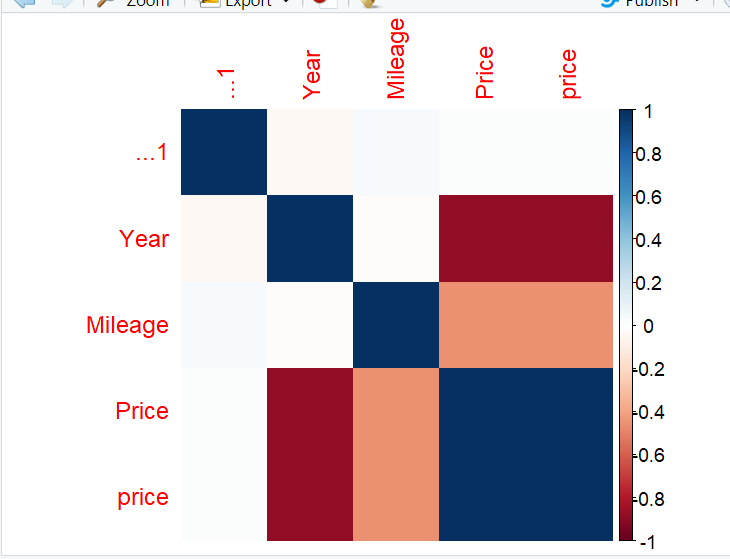
Here, I am working on the 2 graphs to show the relationship with the response variable using scatter plot.



* Here, I used scatter plots and compared the relationship between mileage vs price-target variable, year vs price.
* Graph 1- Scatter plot for mileage vs price is x axis stands for mileage and y axis for price , the slope is falling down from right to left which means more the increase in mileage , lesser the price value.
* Graph 2- Scatter plot for year vs price x axis stands for year and y axis stands for price, which means for older cars , the price is high and for new cars , price is low. This is due to other factors influence also. That is the reason I got graph like this.

1. **Show a correlation plot for numeric values-**

* Each cell explains the relationship between variables. Also, strength and direction of the correlation.
* In this there are few colors, blue means positive correlation, red means negative correlation.
* Light color shades means weak correlation
* It shows a visual portray of the numeric variables helps you to identify the patterns in my dataset.

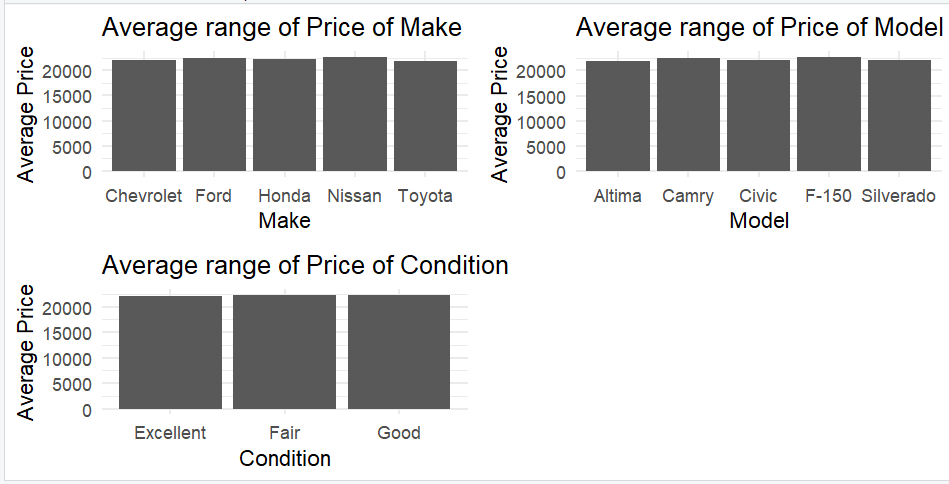


1. **Breakdown of categorical variables**- There are 3 categorical variables- Make, Model, Condition
2. Average value of response for those categories (if response variable is numeric)-

Aggregated the to graph.

**Example**- If graph the price of each model of car and the values are like $1,000,000 for the Altima. It does not make any sense. So we take the average price for each model of car. So Average price of a Altima is $18,000. That is a much more reasonable price to expect if you are buying a new car.  
so, Getting the average is very important .

1. I had plotted the column chart of average Price of Make, Model and Condition putting x as categorical variables and y as price.



1. **Importance of Variables**-

Mileage and year plays an important role in every model of mine. Let me explain.

**Mileage**- We know that this measures the miles drove by the driver. This is a significant predictor value for price variable. It gets the model accuracy.

**Year**- It gets value like in which year the car was introduced and gets the predicted value of price. By this my models clearly explain the significance of these variables.

Now lets go to models.

**Part 2-Model Development and Analysis**

1. **Detail the models you developed and why you chose those models-**

**Classification or regression**-

* + 1. I am doing **regression** for all my three models which are multi linear regression, decision tree and random forest model.
    2. Because in my project, I am using Price as a target variable which is a numeric variable based on other variables. My output will come in numeric forms.

1. **Why were the models you developed best suited to the outcome –**

I developed models to know the relationship between predictor variables and response variable

**Multi Linear Regression**-

*Suitability*- The MLR shows the linear relationship between predictor variables and response variable.

*Selection Reason*-It shows interpretable coefficients that exactly represents the direction relation with predictor and target variable.

**Decision Tree**-

*Suitability*- It shows non linear relationship between predictor and response variables.They can be numerical or categorical.

*Selection Reason*-In my dataset, they represent nonlinear relationship with variables. If there are huge number of variables , they handle a large complex patterns without the use of linearity.

**Random Forest**- It gets together all the decision trees and improve accuracy and decreases overfitting of my model data.

*Selection reason*-It represents the non linear relationship of my variables reducing overfit. If the data is large, it fixes the issue by handling large patterns of data. It is robust and used in every datasets.

1. **Whether the types of data you had, the normalization (or non-normalization) of the data impacted your decision, and why you included the variables-**

Data types- There are 6 variables in my dataset in which 3 are numerical and 3 are categorical variables. Random forest and decision tree have the capacity to handle large data easily , Linear regression requires some work in which it should convert the categorical to numerical to get output.

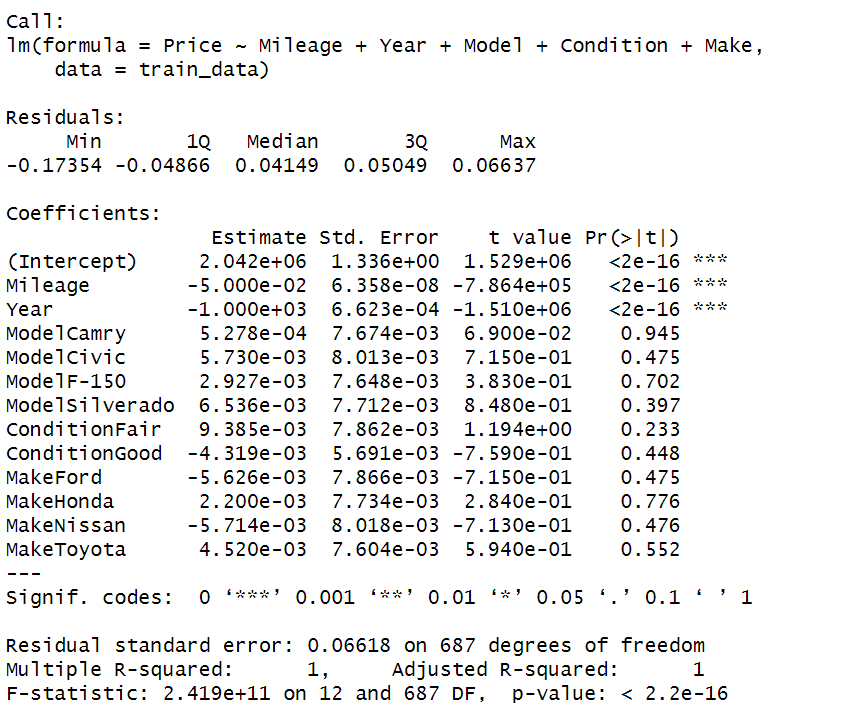
Normalization- In every model, there are different kinds of scaling for predictor variables. For my models, multi linear regression are standing on good scale of variables but other models such as decision tree and random forest have different scales of variables.

Variables Included- Mileage and Year are included for price as they pay attention to the significance in my dataset. It basically comes from predictor variables and response variables which are caught by exploring the analysis of my data to prove a relationship with my variables predictor and response.

1. **Develop and interpret your models-**

I developed three models such as Multilinear regression, decision tree and random forest. Lets start.

1. **Multi linear regression Model-** Here, I explain about the output summary**.**

****

**1. Most Important Variables**- In my model, the most significant codes are Mileage and Year. These are the most important variables.

**2. Impact of outcome-**

**Mileage**: It is a negative coefficient means mileage increases, car price reduces. This is intuitive as higher mileage often indicates more wear and tear on the vehicle.

**Year**: It is a negative coefficient which means old car have low prices compared to new cars.

**3. overall effectiveness of the model-**

**MAE**-0.05795505- It is accuracy value of the actual and predicted prices.

**RMSE**-0.06459707- It finds out the average difference of the actual and predicted prices

**R-Squared** – 1 It suggests a **good** fit for a model

**Residuals**:

min residual is -0.17354, This describes difference between actual and predicted prices is -0.17354.

first quartile residuals is -0.04866 explains first 25 percent residuals "<=" -0.04866.

middle residual 0.04149 describes 50 percent are below.

third quartile is 0.05049 states 75percent residuals "<=" to 0.05049.

max residual is 0.06637 shows bigger difference between actual and predicted prices is 0.06637.

**Coefficients**:

Mileage is -5.000e-02 - This describes each unit increase in mileage,car price reduces putting other variables constant

Year is -1.000e+03 -This indicates as year increases , the price reduces to 1000 dollars putting other variables constant.

Coefficients for categorical variables shows change in price

**For example:**

Lets take model coefficient , ModelCamry is 5.278e-04.This shows that cars have high price compare to other models. It is not statistically significant.

Lets take condition coefficient, ConditionFair is 9.385e-03. This shows that cars having fair condition high price compare to other cars. It is not statistically significant.

**Residual standard error:**

It is having 0.06618. This shows actual prices goes from predicted prices to 0.06618 dollar

**Multiple R-squared and Adjusted R-squared:**

Multiple R-squared and Adjusted R-squared values are 1- This shows variables are predicted in price of cars which is excellent fit for my model.

**F-statistic:**

F-statistic value is 2.419e+11 which is large value.

p-value < 2.2e-16 shows highly significant.

(IT IS **GOOD** BECAUSE OF LOW MAE, RMSE VALUES AND HIGH R SQUARED VALUES)

1. **Decision Tree Model**- Here, I explain about the output summary.

**Most Important variables-** Here, I see mileage and year plays an important role in predicting the trees outputs.

**For example-** In below chart, Year is greater than 16 , means mileage is greater than 70,000 miles with year greater than equal to 13 gives price 29,000 dollars

**Impact of outcome-**

Year and Mileage are extreme significant scores. It means they are strongly good to predict prices.

**Example**-Cars with low mileage have a high prices.

**overall effectiveness of the model-**

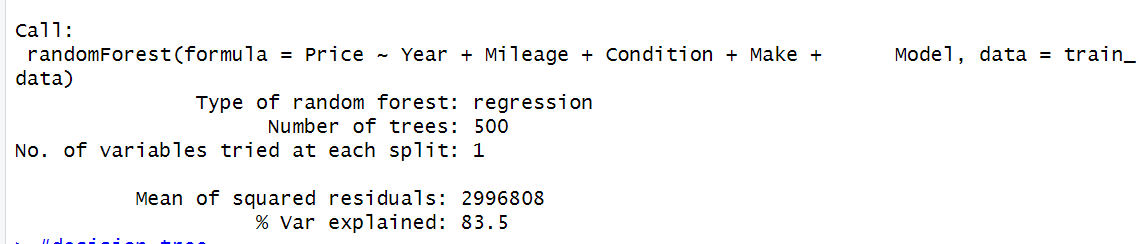
There are 500 trees. Here, I found the Mean absolute error , Root mean squared error and R squared.

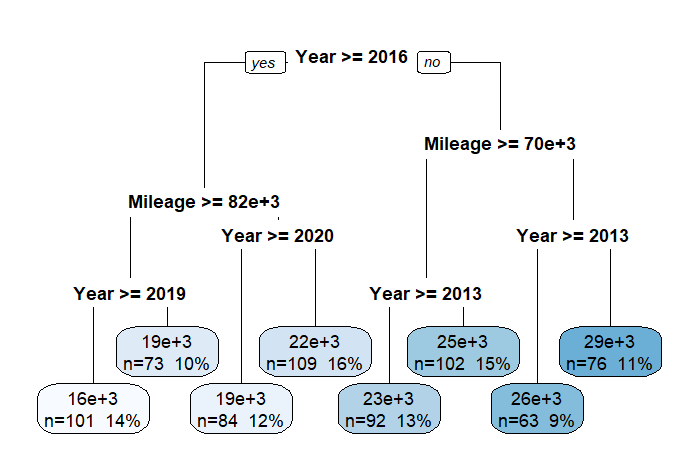
**MAE**- 1074.508 - It is accuracy value of the actual and predicted prices.

**RMSE**- 1324.87 - It finds average difference of the actual and predicted prices.

**R-Squared** - 0.9004521 -It suggests a fit for a model

(IT IS BAD BECAUSE OF HIGH MAE, RMSE VALUES AND LOW R SQUARED VALUES)





1. **Random Forest Model-**

**Most important variables**- Mileage and year are important variables as mileage have 2300679079 and year have 6262461114 Which are high significant.

**Impact of outcome**-As Year and Mileage are important scores. This means that latest new cars with lower mileage have higher prices.

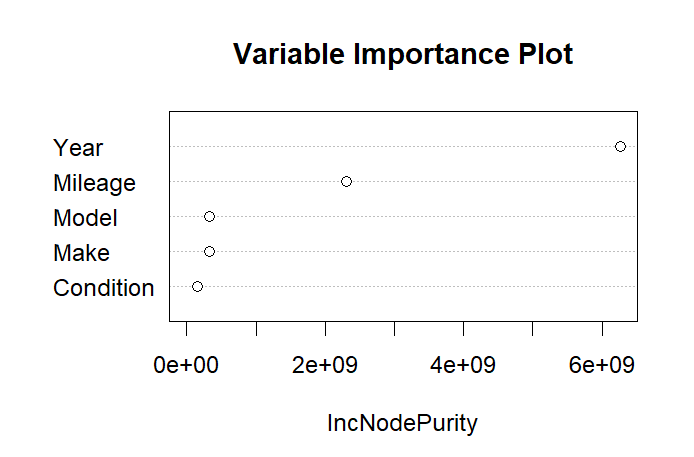
**Overall effectiveness of model**-

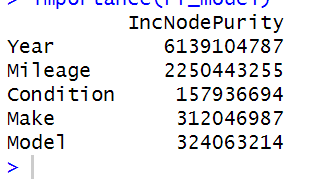
**MAE**: 1380.475 - It is accuracy value of the actual and predicted prices.

**RMSE**: 1716.742- It finds out the average difference of the actual and predicted price.

**R-squared**: 0.832854- It suggests a fit for a model

(IT IS BAD BECAUSE OF HIGH MAE, RMSE VALUES AND LOW R SQUARED VALUES)





1. **Comparisons of models**-

Here, I had compared the models.

* I see that that decision tree and random forest MAE and RMSE values are low compared to MLR model.
* R squared values are lesser than MLR model.
* MLR have low MAE and RMSE values with High R squared values.

Overall, The Multi Linear Regression model is the best model.

**6.Model Comparisons**-

Compare Models-

Model           MAE          RMSE R\_squared

1 MLR     0.05795505 0.06459707 1.0000000

2 Decision tree 1074.50799574 1324.86951064 0.9004521

3 Random forest 1380.47541094 1716.74198825 0.8328540

* I see that that decision tree and random forest MAE and RMSE values are low compared to MLR model.
* R squared values are lesser than MLR model.
* MLR have low MAE and RMSE values with High R squared values.

Overall, The Multi Linear Regression model is the best model.

**7.Proof – (Charts and statistics)**

**Charts**- I compared the values of MAE, RMSE and R squared with graphs and portrayed.

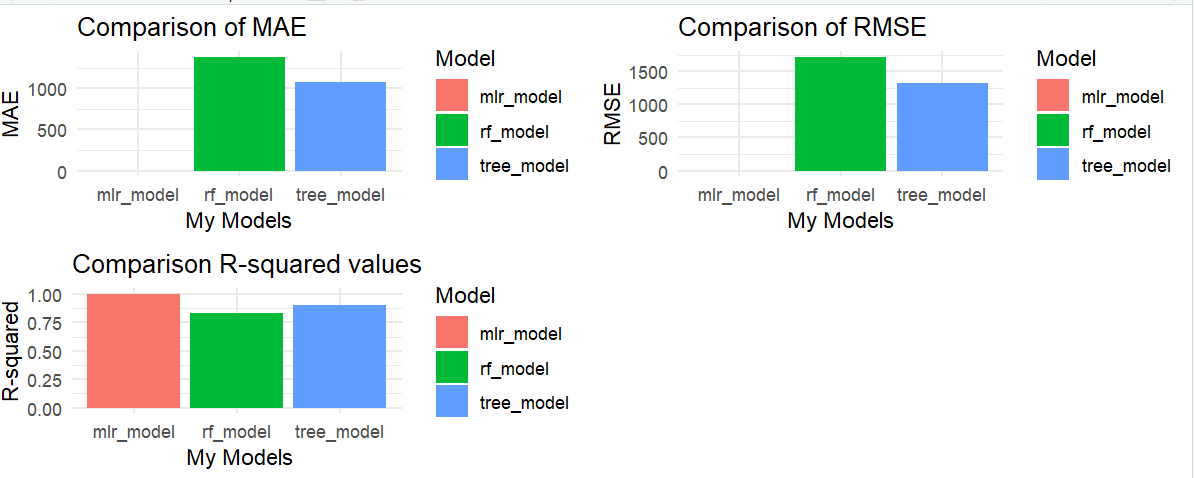
**Statistics**-

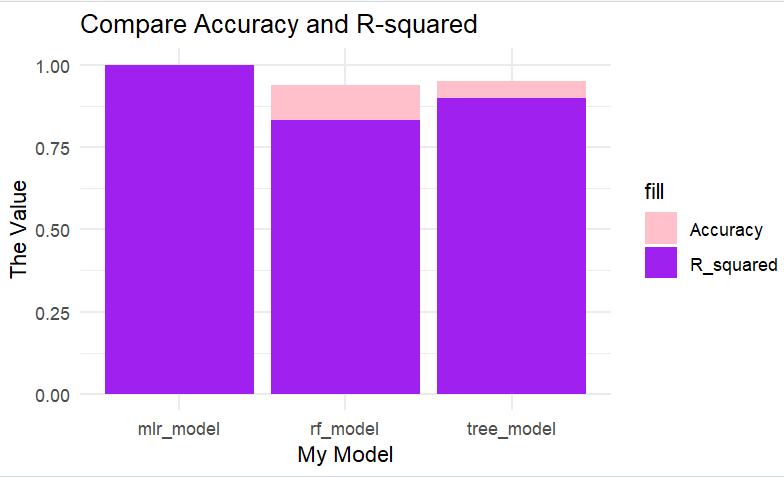
Model           MAE           RMSE R\_squared

1 MLR     0.05795505 0.06459707 1.0000000

2 Decision tree 1074.50799574 1324.86951064 0.9004521

3 Random forest 1380.47541094 1716.74198825 0.8328540





1. **Why do you think that your best model performed better than the others?**

My Model Multi Linear Regression performed better because of many factors. It gave the access to each variable in my dataset to understand the response variable easily. It also have a sophisticated strong relation between my response and other variables than compared to other models. As we know that overfitting leads to poor outputs. In my case, MLR has done a good job compared to other models with R squared value of 1.

Also, my dataset is small. It did not trouble me and it helped models to work easily.

PART 3- CONCLUSION

1. **Reiterate findings-**

a) My purpose of my project is to predict the price against all other variables such as make, model, condition , mileage , year using multi linear regression, decision tree and random forest models.

b) My outcome was comparing the actual price vs predicted price and I successfully did with MLR model by comparing metric values such as mae, rmse and r squared values which are perfect fit

c) My best model was Multi Linear Regression model compared with other models. It has high r squared value and low mae and rmse values. It predicted very good accurate prices of the car.

d) The key variables were mileage, make, model, condition ,year. The important variables were mileage and year. They captured a strong linear relationship between response and predictor variables which gave accurate prices.

2. **Why you believe that you answered the question of your research topic-**

Dataset good or bad in making predictions-

Ofcourse, my dataset was good in making predictions because it gave a very good accuracy in making good prediction of my price. I had attached the graph accuracy above in chart section. The performance was very good because it worked relation between predictor and response variable.

Final observations-

Based on research, methodology and the process my car predictor variables helped my response variable to get a accurate price outcome and utilized the machine learning algorithms. The importance of model and comparison gave a predicted output.

3. **Next steps**-

a. I think for future research it would be better if added extra additional variables or features. Such as location of the car and market trends. For example- If I don’t know where to buy the car and don’t know which car is leading in market it becomes easy for me. So, I would like to expand these features.

b. Interpretation of models sometimes becomes difficult and challenging and yes I strongly acknowledge that the models give a good predictive accuracy but am talking about the interpretation. Understanding the concept of interpretation makes life easy with clear outputs.

4. **What you have learned-**

There are many things that I have learnt from this project. Let me explain.

Data cleaning is very very important when I build a model. It is a root for any model and coming to building the model, I feel each model have different styles to know the relationships and the accuracy results. I literally found difficulty working on models but I practiced more and more until I make it.

Always being focused and maintaining it is very important. For example-If am done with R, I should still be in touch with concepts. This is the one thing I will take to future classes.

For every person, there should be a strong will power and interest on the areas. When I was in India, I always had a strong will power to become a data analyst. I decided to do masters.

But before that, I started doing some courses in india. I used to go to small tuitions and learn the basics of R, etc.

It is easy to register for programming classes in the university. But before registering I should know how far I can take this to or what extent I can make it successful. Am sure I will change graduate programs.

Last by not least, I want to become a Pro Monkey whatever I study. Not just a beginner monkey or middle monkey. Thanks